www.bsc.es

Madrid, 29 February 2016

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Big Data for climate and air quality

Francisco J. Doblas-Reyes
BSC Earth Sciences Department

iCrea

IC3

BSC Earth Sciences Department

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

What

Environmental forecasting

Why

Our strength …
- … research …
- … operations …
- … services …
- … high resolution …

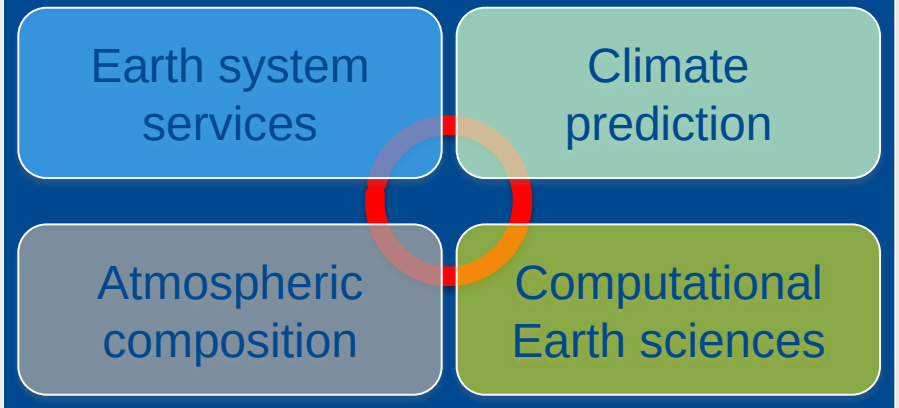
How

Develop a capability to model air quality processes from urban to global and the impacts on weather, health and ecosystems

Implement climate prediction system for subseasonal-to-decadal climate prediction

Develop user-oriented services that favour both technology transfer and adaptation

Use cutting-edge HPC and Big Data technologies for the efficiency and user-friendliness of Earth system models

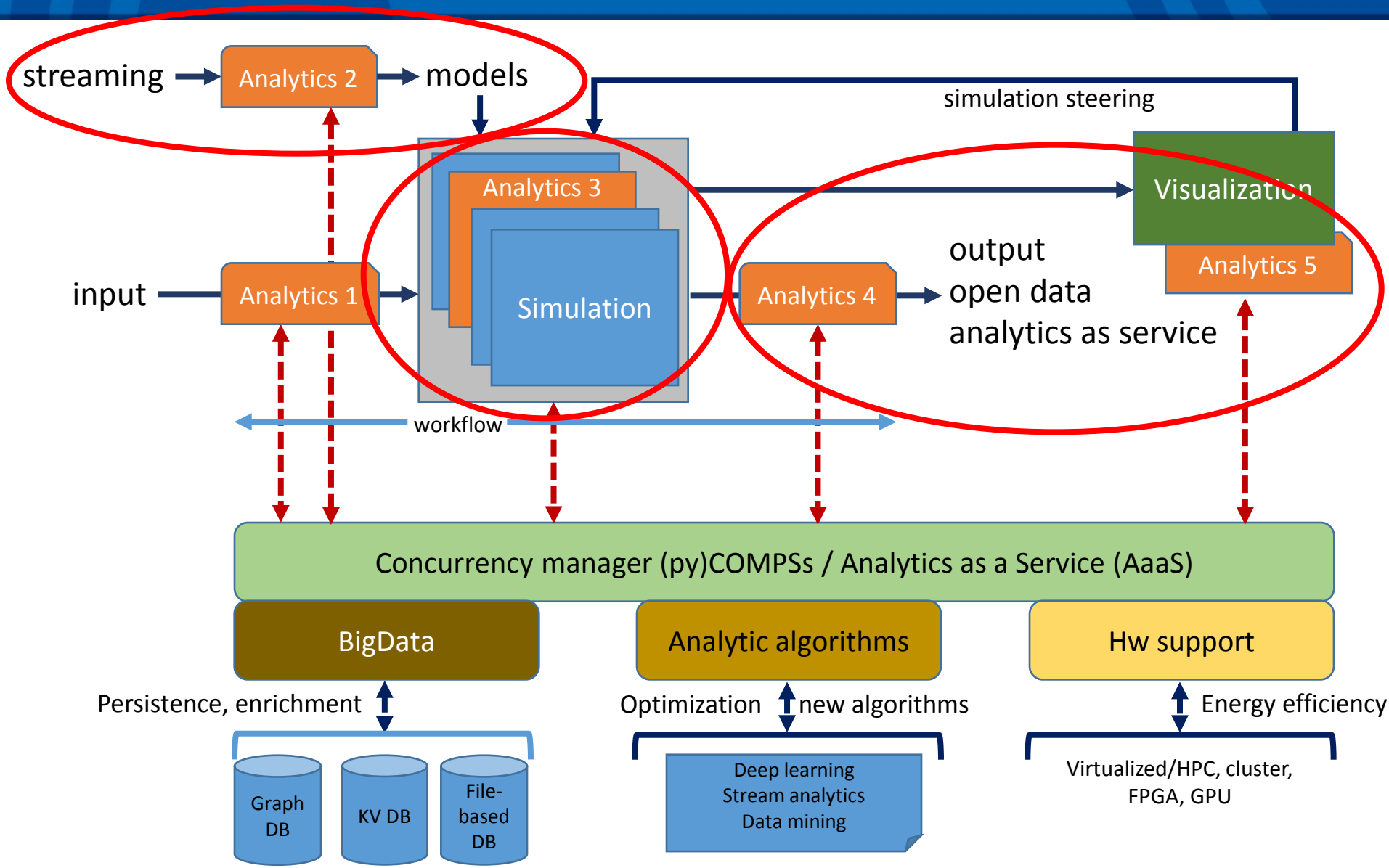


Big Data in our field

- There are problems involving large, complex datasets: operational weather and air quality prediction.
- There are large problems involving data: simulation of anthropogenic climate change.
- And there are Big Data problems: dealing with heterogeneous data sources to produce end-user information with a weather, climate and air quality component.

I will not address the issue of open data and assume that access to data is not a difficulty.

A conceptual model from CS



Case 1: Data streaming for air quality forecasting

Case 2: Simultaneous analytics and HPC in climate prediction

Case 3: Analytics as a service

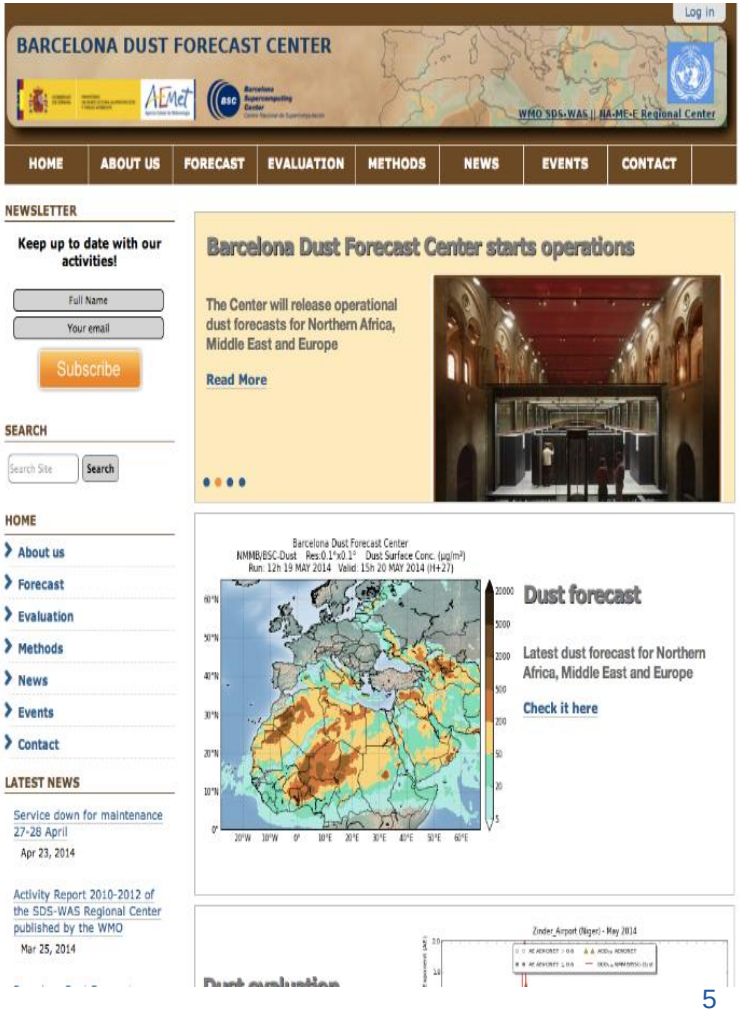
Operational mineral dust forecasts

Barcelona Dust Forecast Center (BDFC) and Sand and Dust Storm-Warning and Advisory System (SDS-WAS) for North Africa, Middle East and Europe, both operated jointly by BSC-CNS and AEMET:

- BDFC is the first specialized WMO centre for mineral dust prediction, and provides forecasts to WMO GTS, EumetCAST and AEMET.
- SDS-WAS NAMEE is a research and development multi-model system gathering model output, AERONET observations and satellite data.

http://dust.aemet.es

http://sds-was.aemet.es



CALIOPE air quality operational forecasts

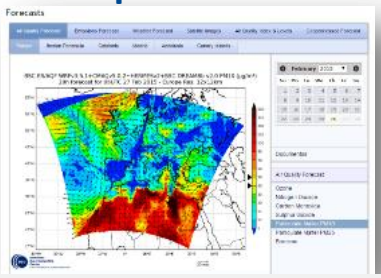
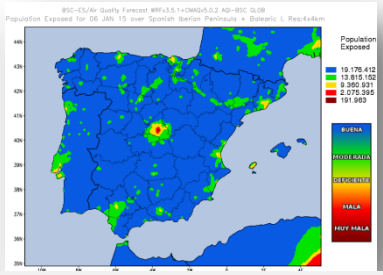
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

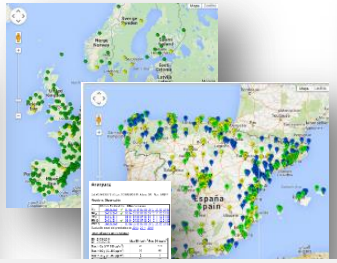
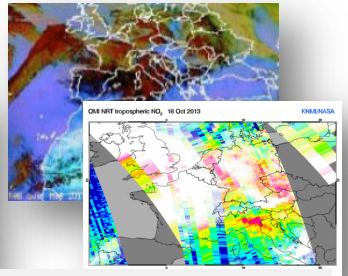
EXCELENCIA SEVERO OCHOA

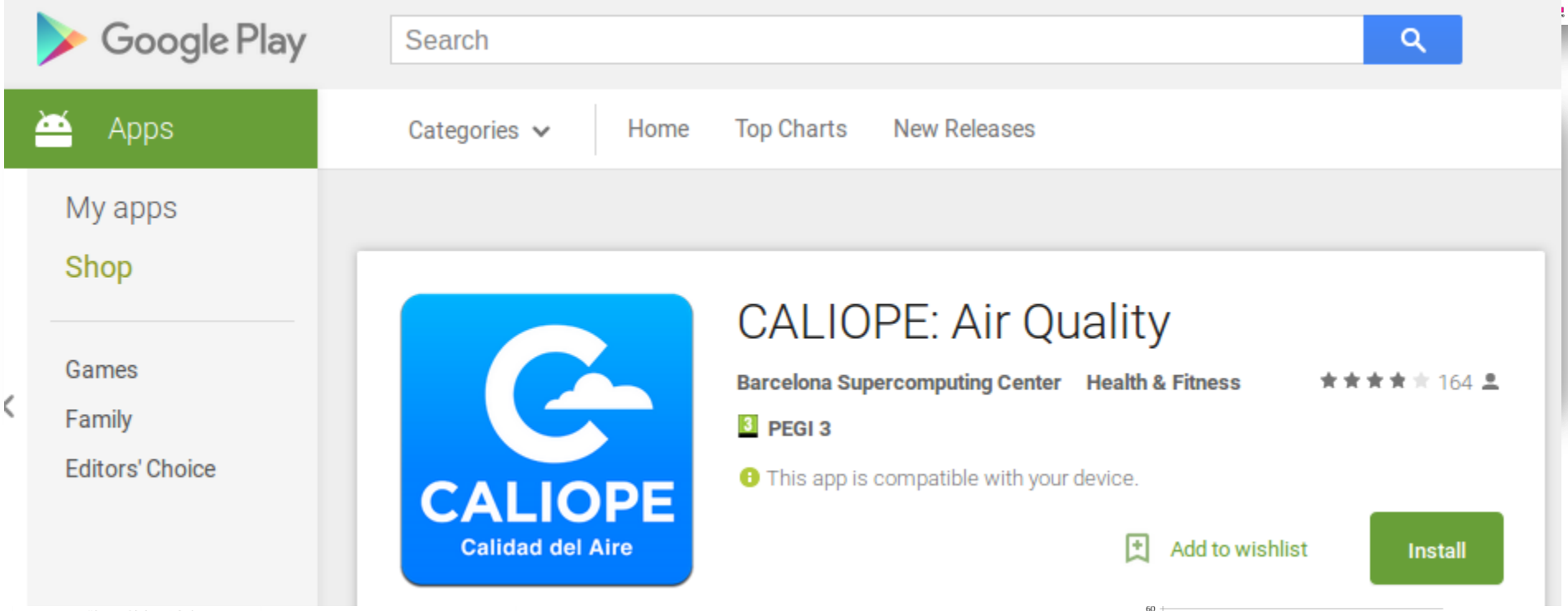
AQF CALIOPE system: daily forecast and evaluation

Daily forecast for **meteorology, emissions and air quality**: Europe (12km), Iberian Peninsula (4km), Andalusia, Catalonia and Madrid (1km), since 2007

Forecast products



PM10_KF annual average skill evolution (2011-2014)

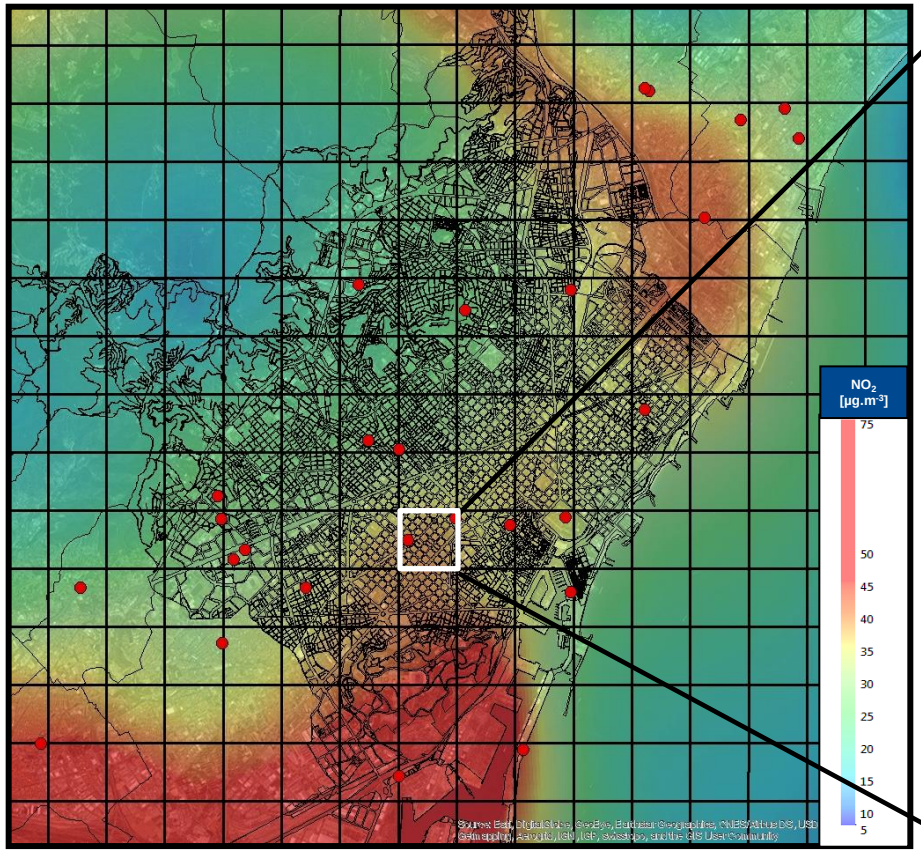
New air quality management tools

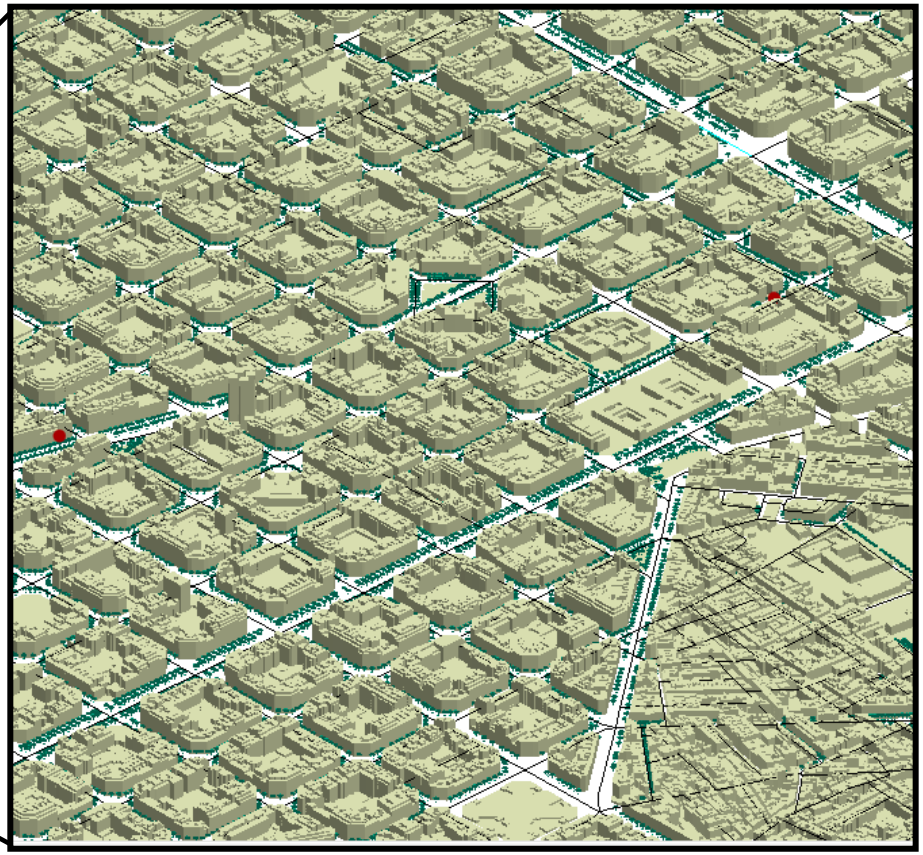
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Where we are now

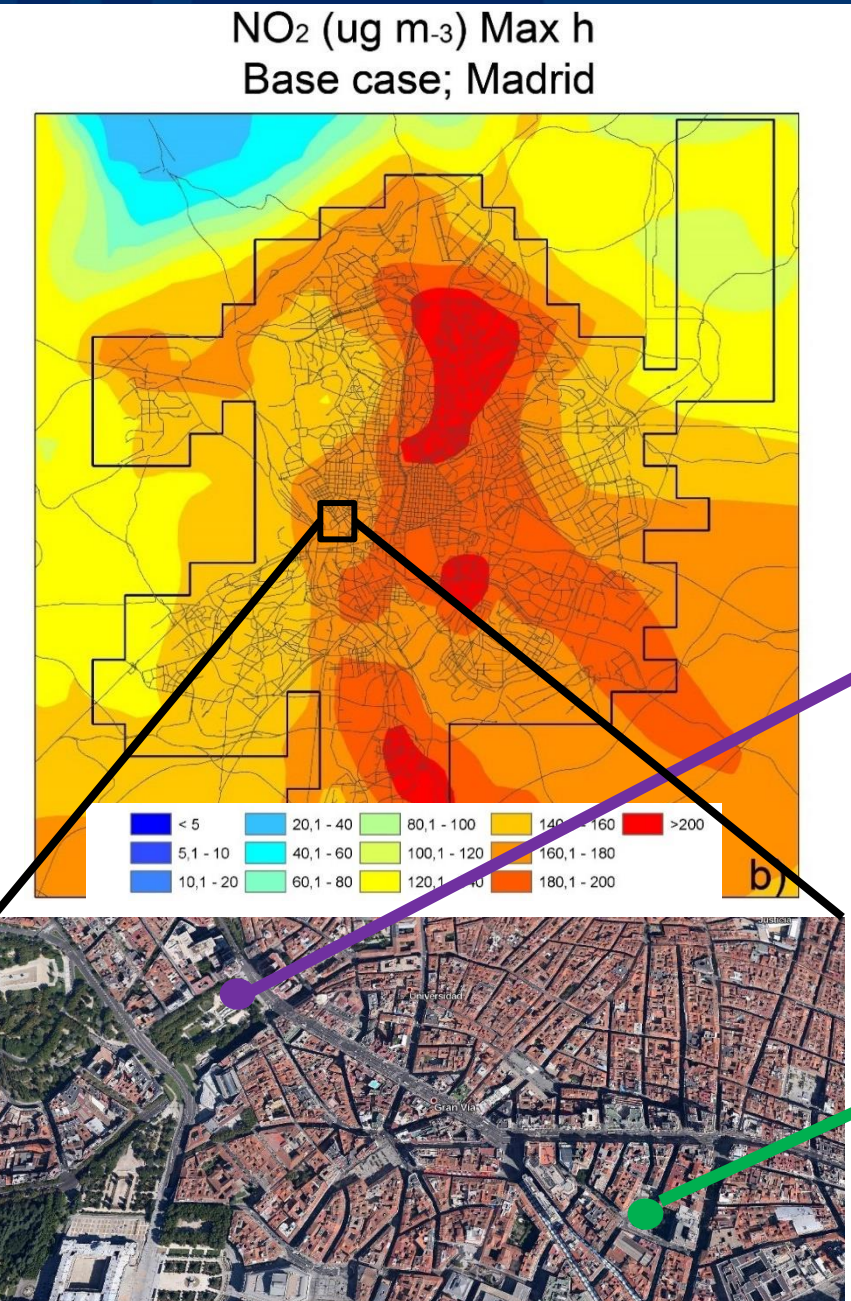
Where we want to be





Objective	To develop an air quality model based on a CFD coupled to an atmospheric chemistry model at city scale, enhanced by the use of Big Data technologies, to assess urban air quality.

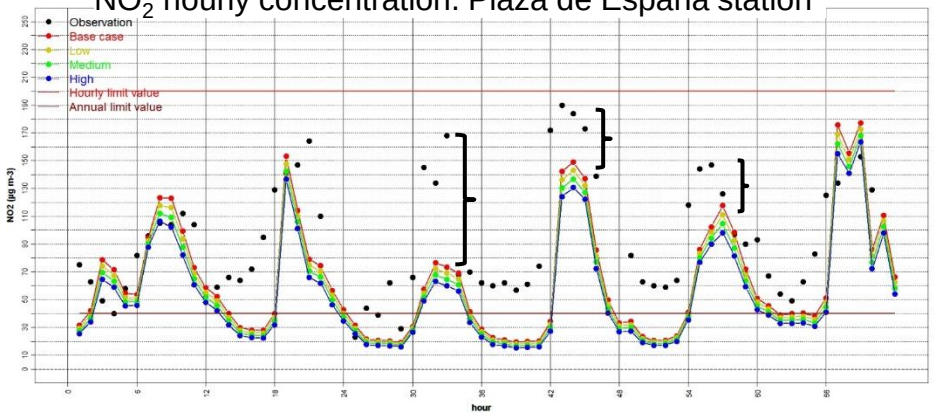
Need of a more detailed view



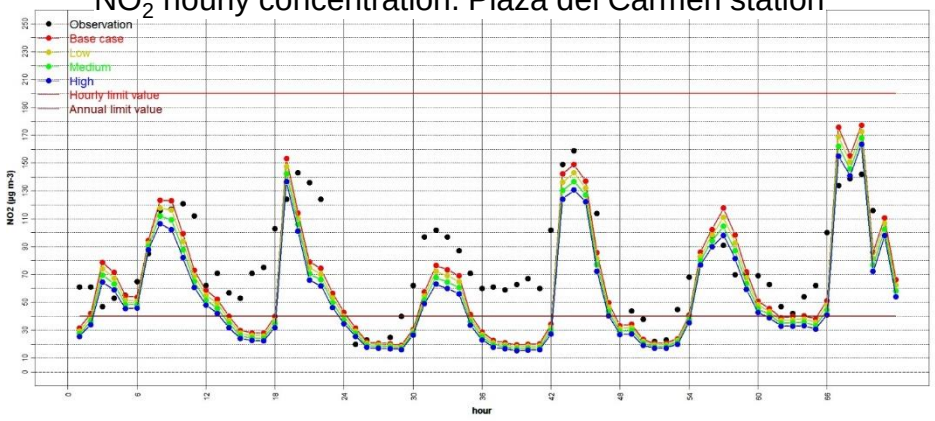
1) Mesoscale models (current solutions) show a satisfactory performance to simulate background concentrations (e.g. Plaza del Carmen).

2) However, they are not able to reproduce street-level strong concentration gradients (e.g. Plaza de España located close to Gran Via); purpose-built tools are needed.

NO_2 hourly concentration. Plaza de España station



NO_2 hourly concentration. Plaza del Carmen station



Data streaming for emissions

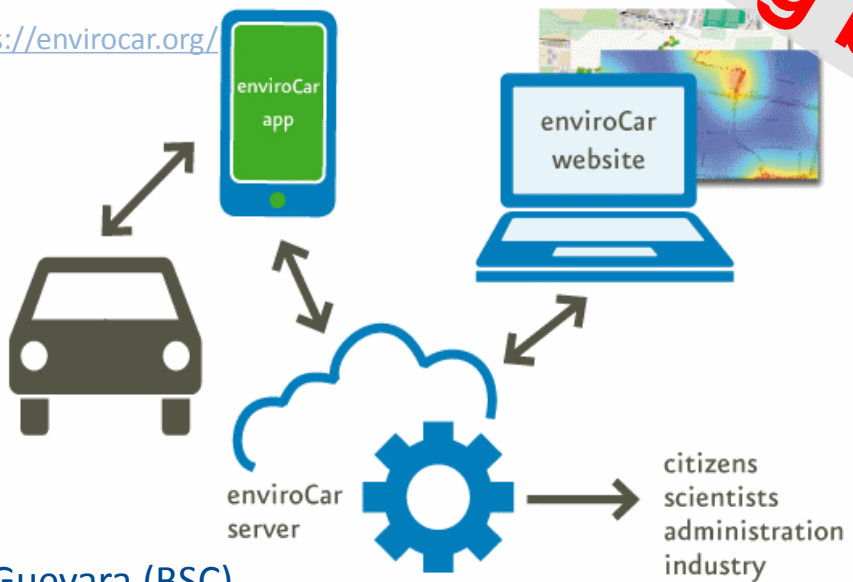
High-resolution air quality modelling requires appropriate emissions

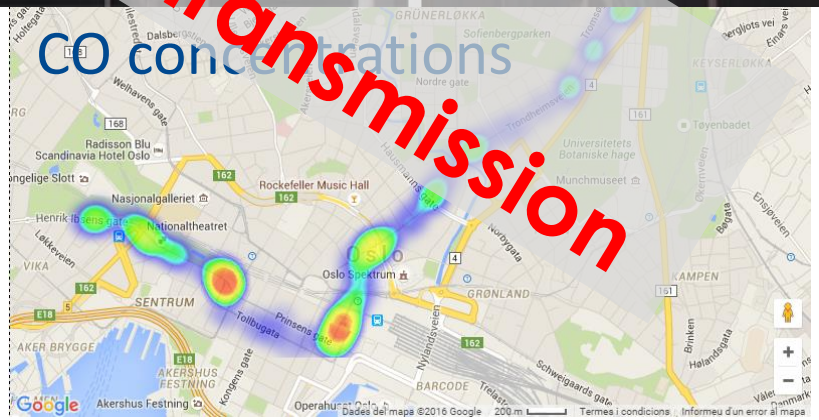
- Collection and processing of sensor-generated data to feed a real-time emission model (and to validate air quality predictions)
- Providing the sensors with the adequate technology is a challenge
- Managing large volumes is another one: sampling 10 Hz, ten variables ~30 MB/day, city-wide ~300 GB/day (10,000 vehicles)

Many more sensors possible

Need for computing before transmission

https://envirocar.org/





M. Guevara (BSC)

Case 1: Data streaming for air quality forecasting

Case 2: Simultaneous analytics and HPC in climate prediction

Case 3: Analytics as a service

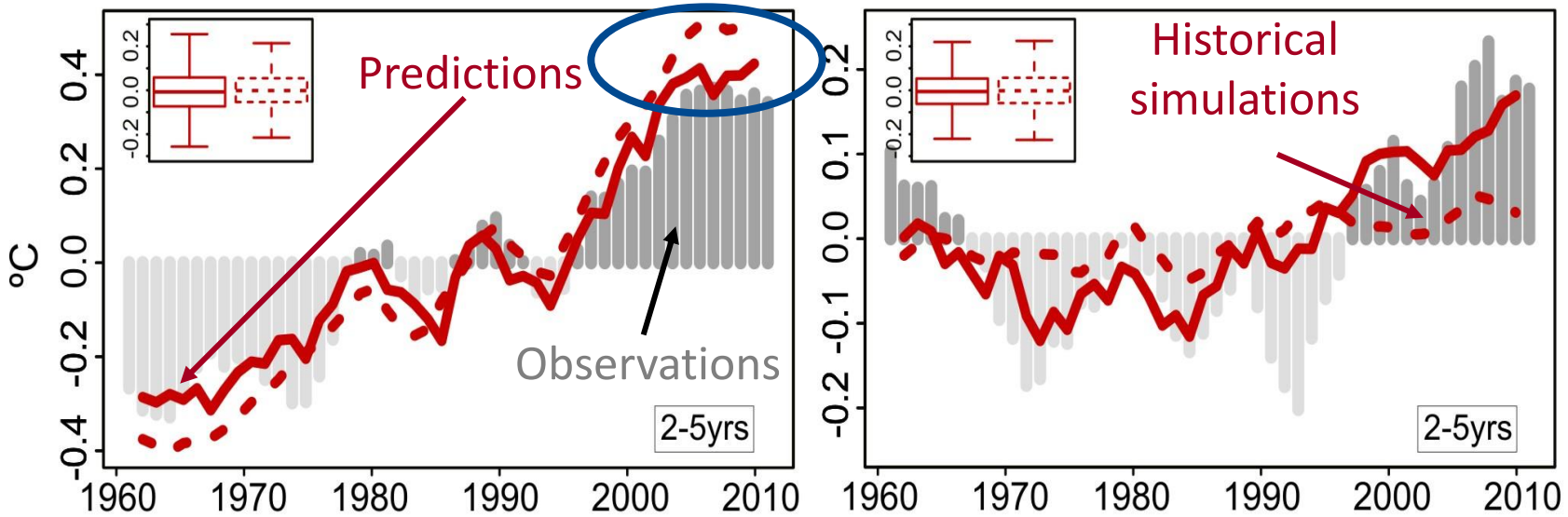
Decadal climate predictions

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA

Global-mean near-surface air temperature and AMV against GHCN/ERSST3b for forecast years 2-5.

Global mean surface air temperature (GMST)

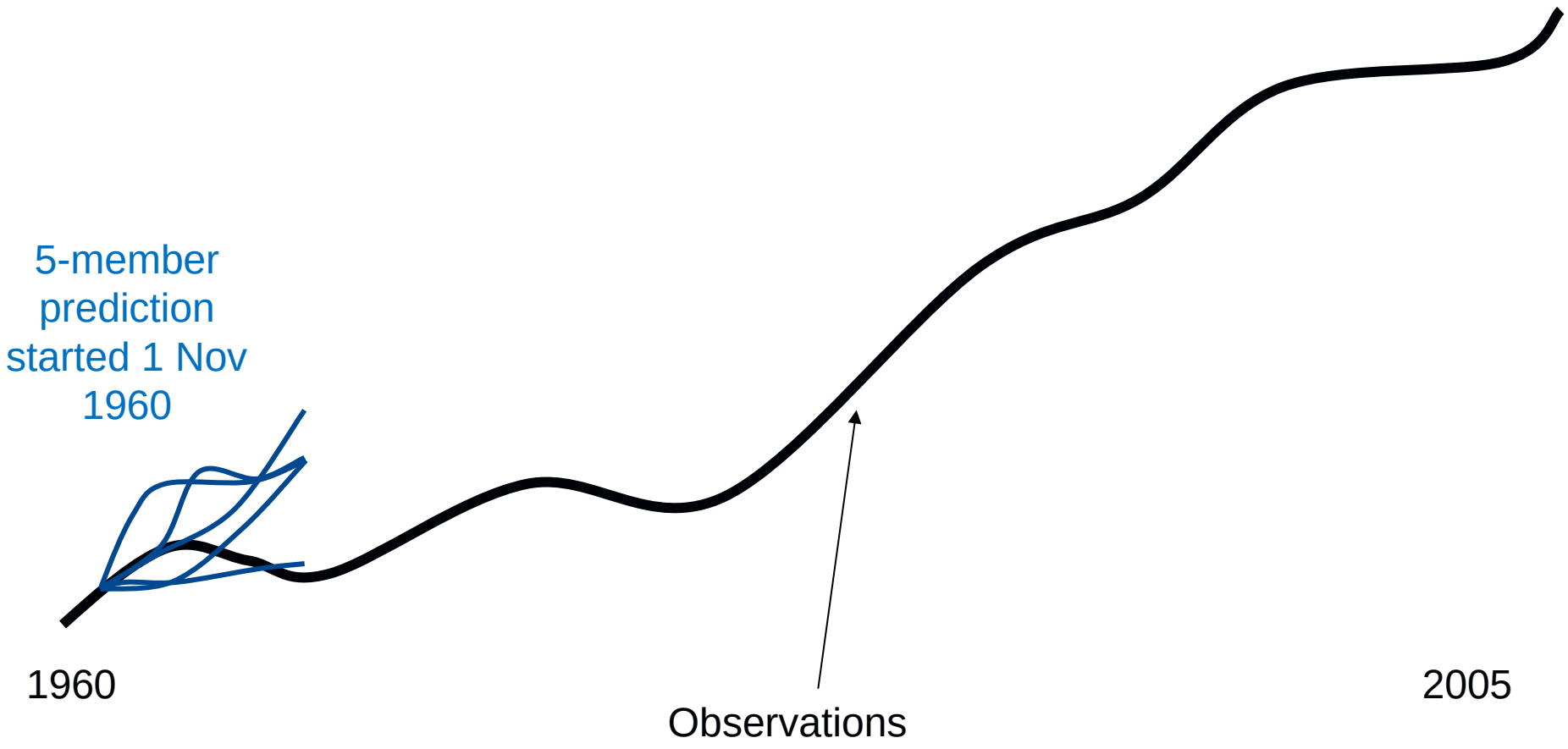
Atlantic multidecadal variability (AMV)



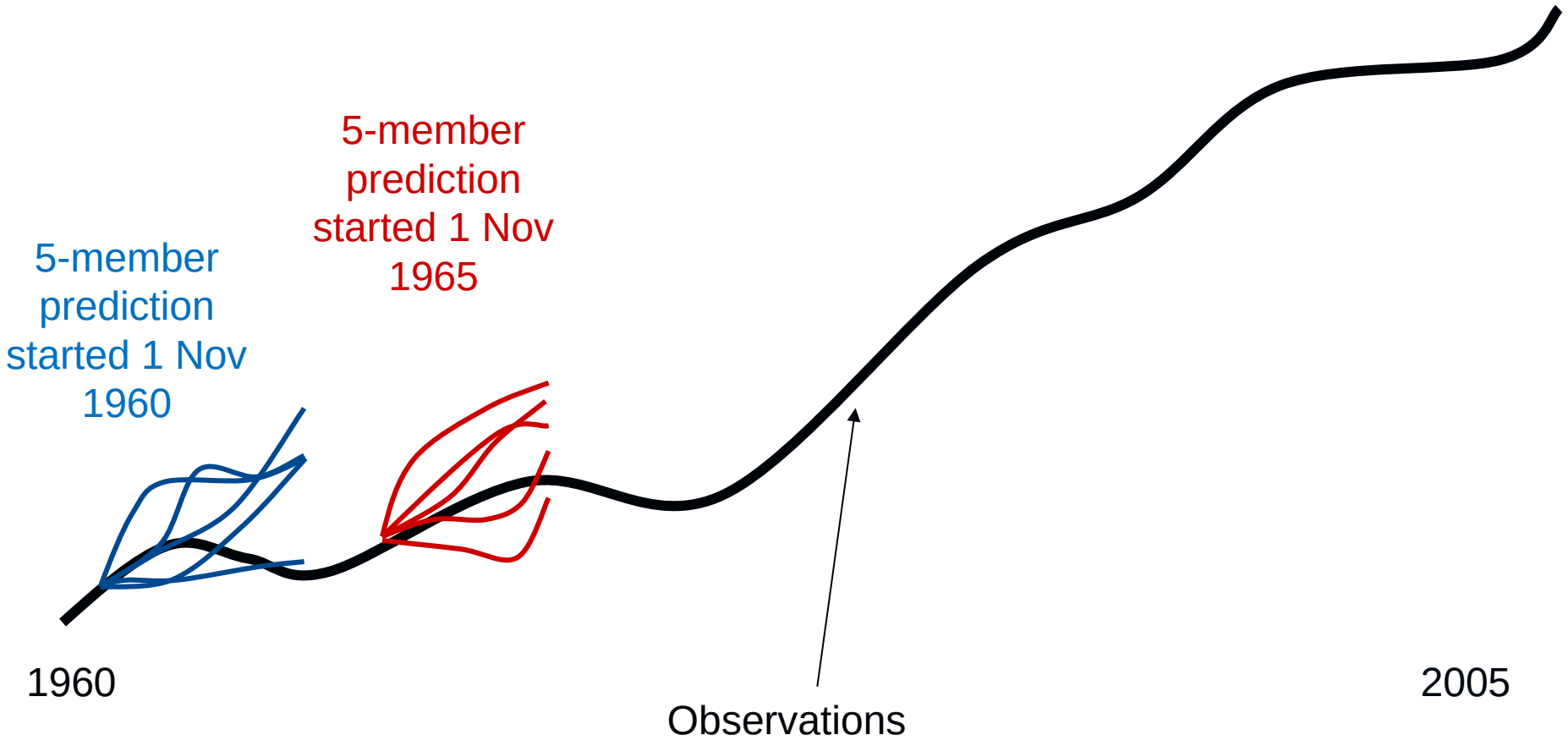
Initialised simulations reproduce the global temperature and some of the AMV tendencies and suggest that initialization corrects the forced model response and phases in internal variability.

Doblas-Reyes et al. (2013, Nat. Comms.)

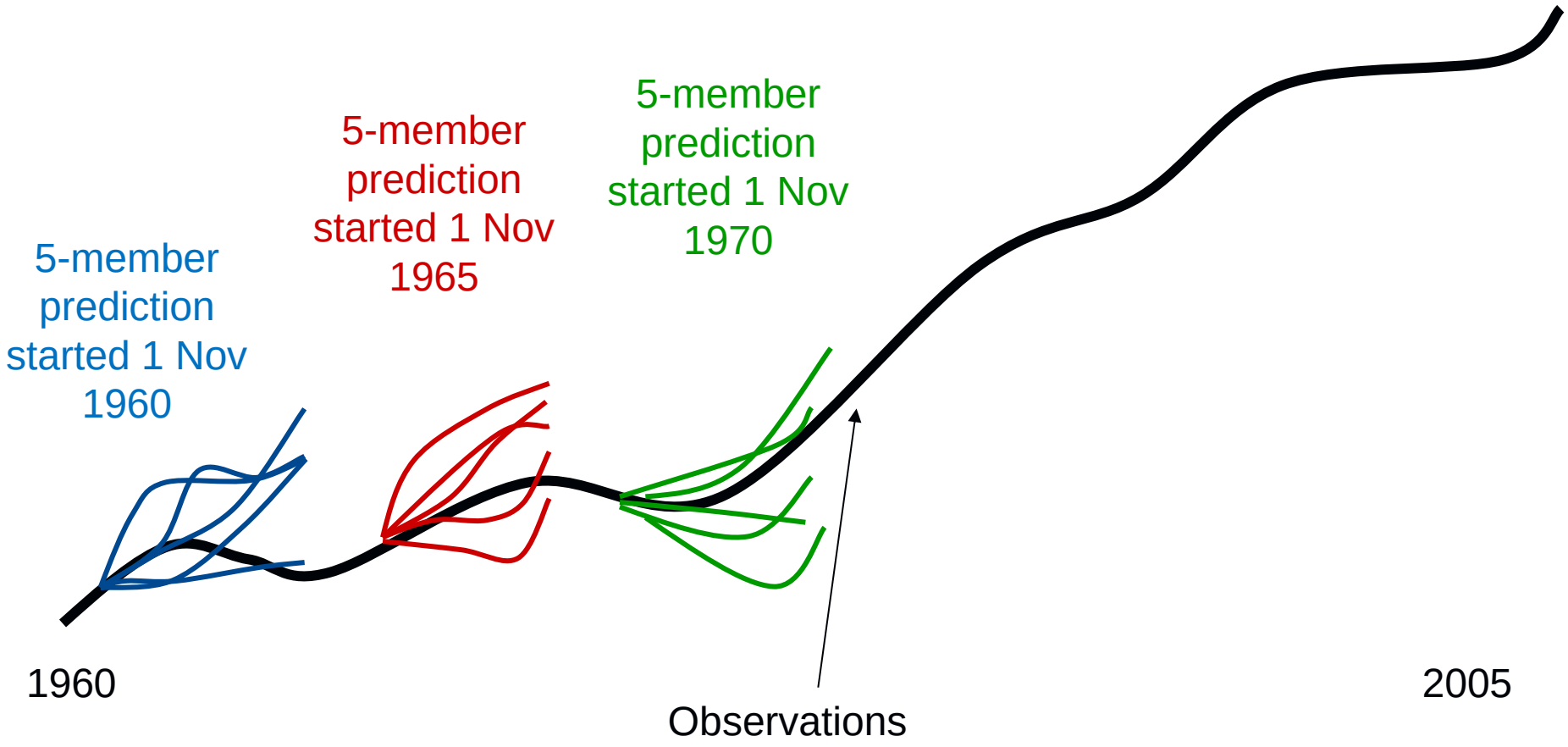
Climate predictions



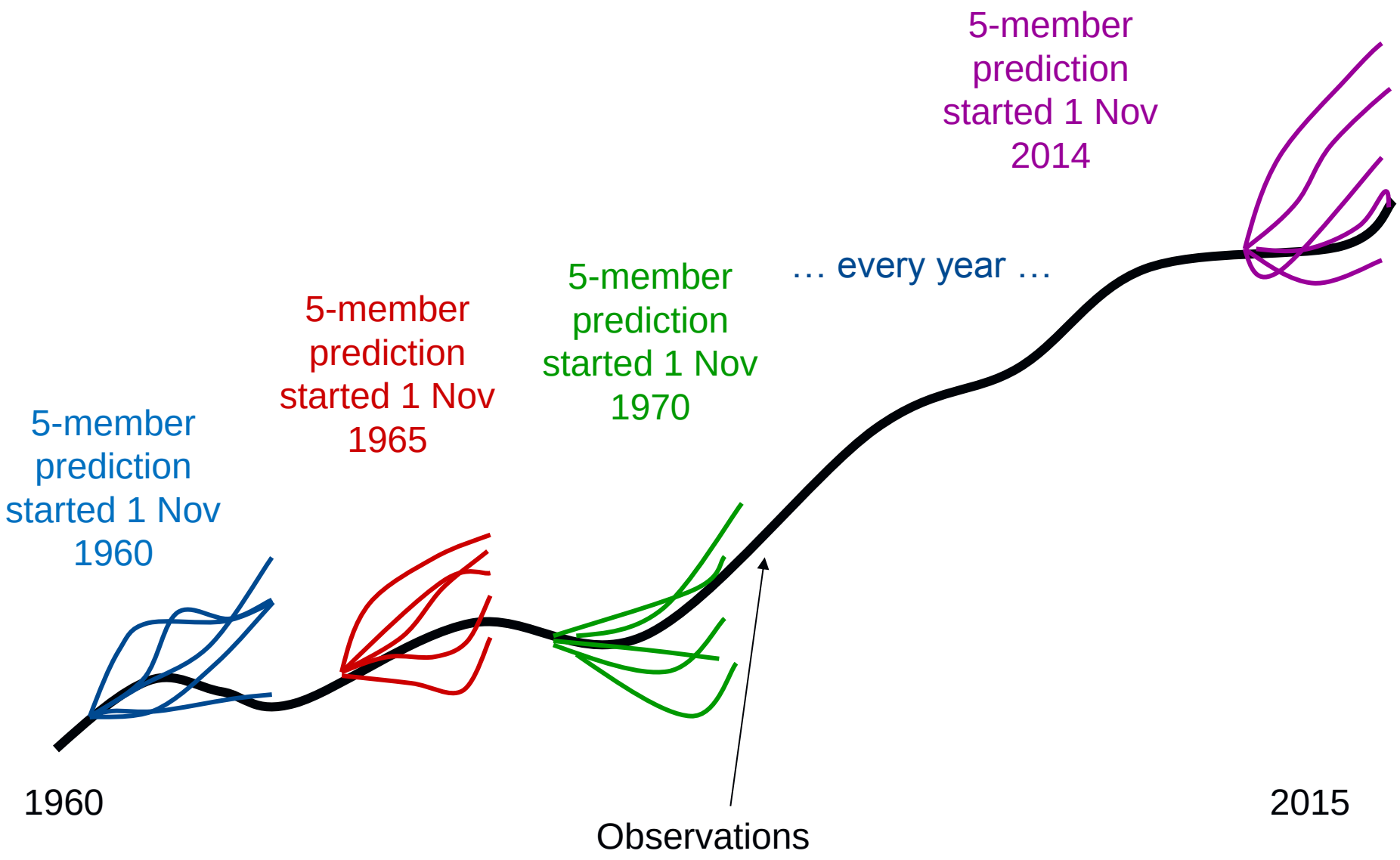
Climate predictions



Climate predictions

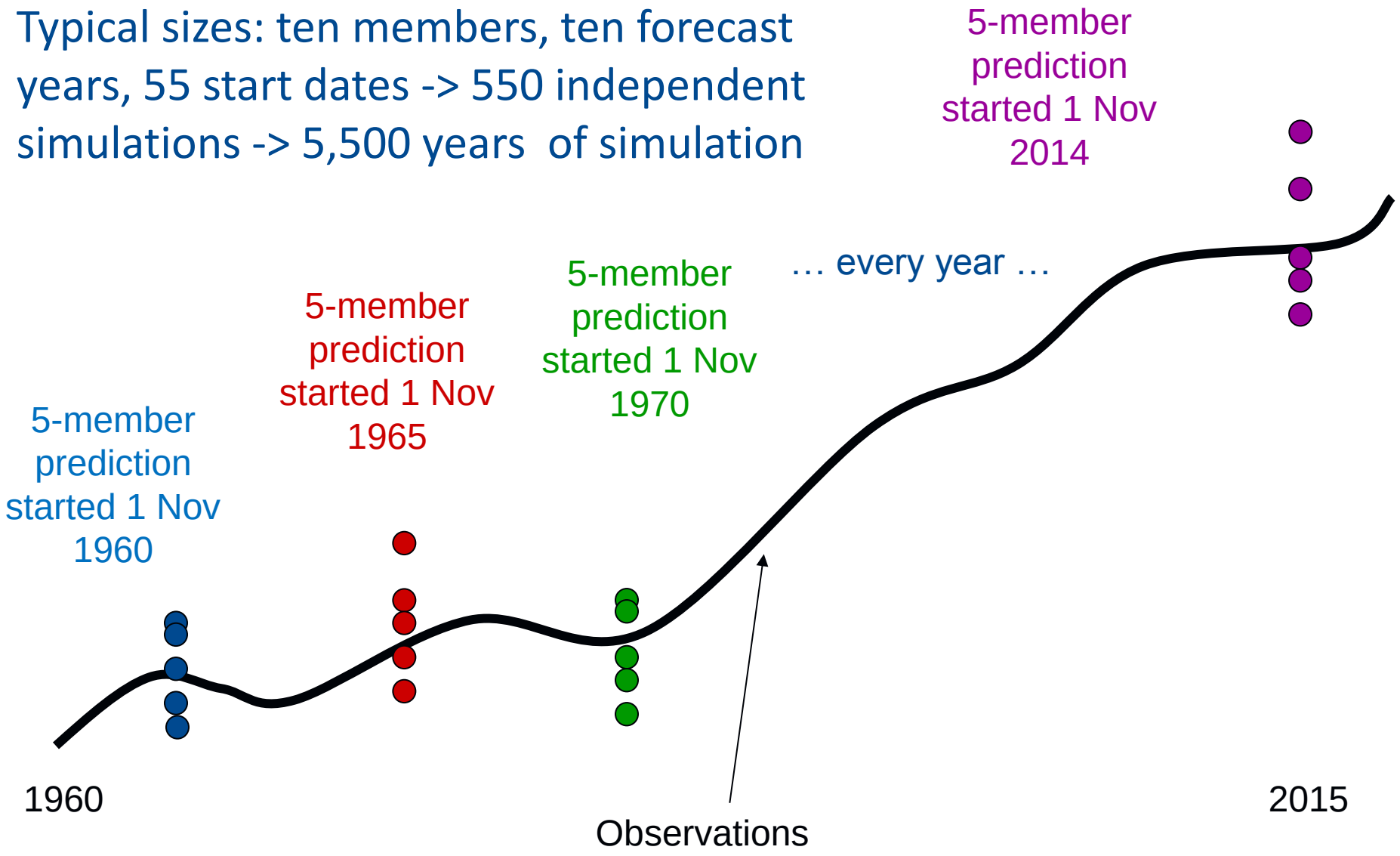


Climate predictions



Climate predictions

Typical sizes: ten members, ten forecast years, 55 start dates -> 550 independent simulations -> 5,500 years of simulation



Running global climate predictions

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA

Climate prediction allows running jobs independently and simultaneously by wrapping together ensemble members for different start dates. This is not trivial parallelisation.

A workflow manager is required.

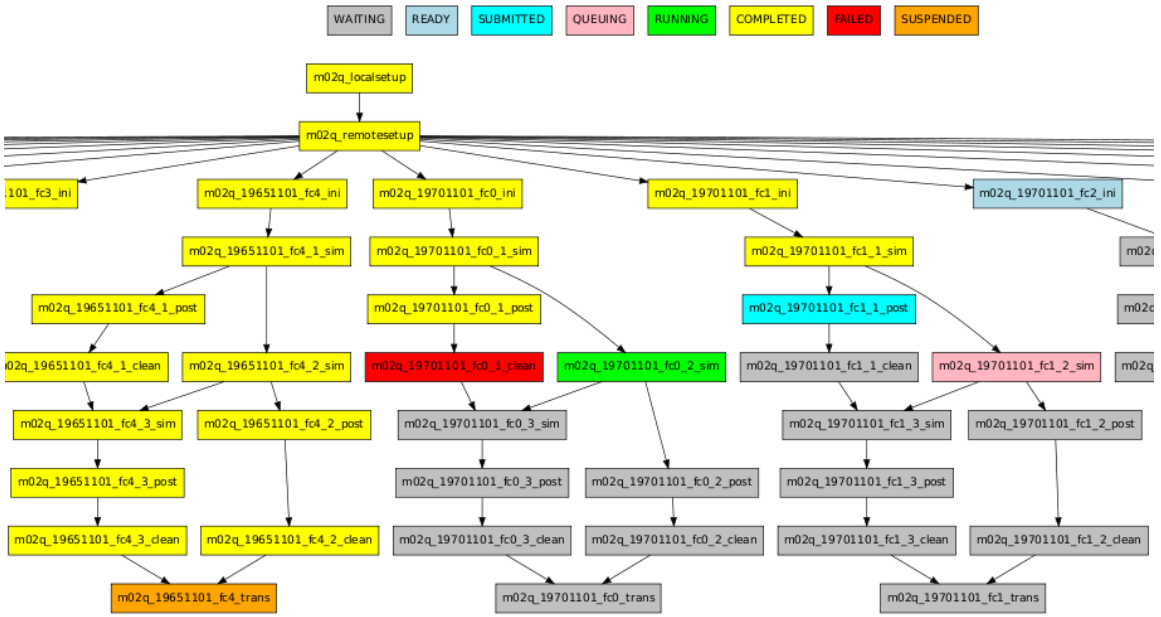
5,500 simulated years -> 1.4 PB

EC-Earth 3 at Lindgren, PDC						
Number of Start Dates		1	5	10	10	20
Number of Members		1	5	5	10	10
Number of Independent Simulations		1	25	50	100	200
T159-ORCA1	Cores	144	3600	7200	14400	28800
	Wall-clock Time (Hours) / year	5	5	5	5	5
	CPU Time (Hours) / year	720	18000	36000	72000	144000
	Output Size (GB) / year	10,80	480	960	1920	3840
T255-ORCA1	Cores	360	9000	18000	36000	72000
	Wall-clock Time (Hours) / year	5	5	5	5	5
	CPU Time (Hours) / year	1800	45000	90000	180000	360000
	Output Size (GB) / year	19,20	5184	10368	20736	41472
T799-ORCA025	Cores	1104	27600	55200	110400	220800
	Wall-clock Time (Hours) / year	40	40	40	40	40
	CPU Time (Hours) / year	44160	1104000	2208000	4416000	8832000
	Output Size (GB) / year	256,80	6420	12840	25680	51360

Workflows: Autosubmit

- **Automatisation:** Preparing and running, postprocessing and output transfer, all managed by Autosubmit. No user intervention needed.
- **Provenance:** Assigns unique identifiers to each experiment and stores information about model version, configuration options, etc
- **Failure tolerance:** Automatic retrials and ability to repeat tasks in case of corrupted or missing data.
- **Versatility:** Currently runs EC-Earth. NEMO and NMMB models **on several platforms**.

Workflow of an experiment monitored with Autosubmit (yellow = completed, green = running, red = failed, …)

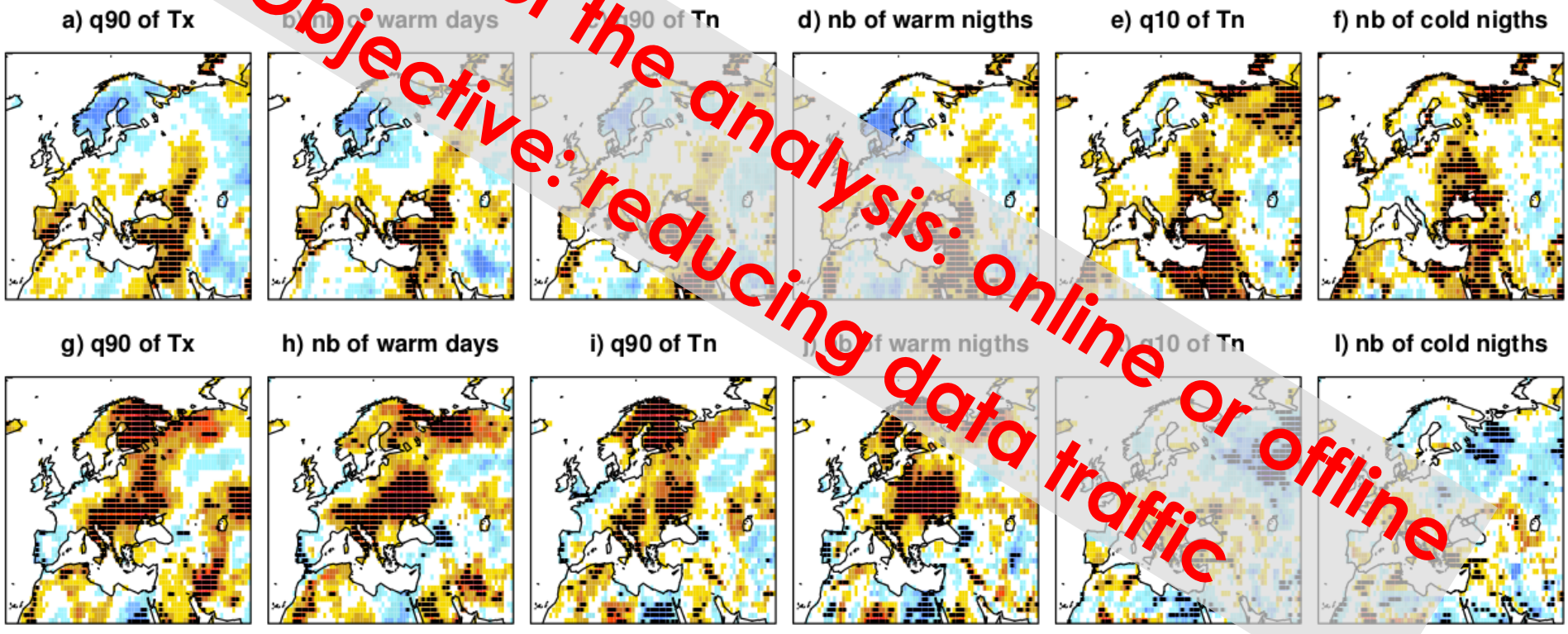


Predicting extremes

JJA near-surface temperature correlation of the ensemble mean from experiments with a climatological (top) and difference with one with realistic (bottom) land-surface initialisation. Results for EC-Earth2.3 started in May over 1979-2010.

Two ways for the analysis: online or offline

Objective: reducing data traffic



Prodhomme et al. (2015, Clim. Dyn.)

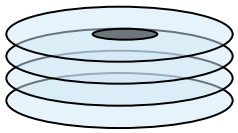
Efficient data analysis

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA

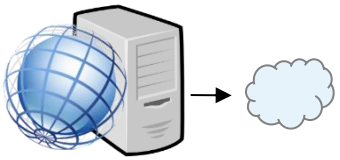
S2dverification is an R package to verify seasonal-to-decadal forecasts by comparing simulations with observational data. It allows analysing data available either locally or remotely. **It can also be used online as the model runs.**

The user demands both APIs and portals

s2dverification package



LOCAL STORAGE



ESGF NODE
or
OPeNDAP SERVER

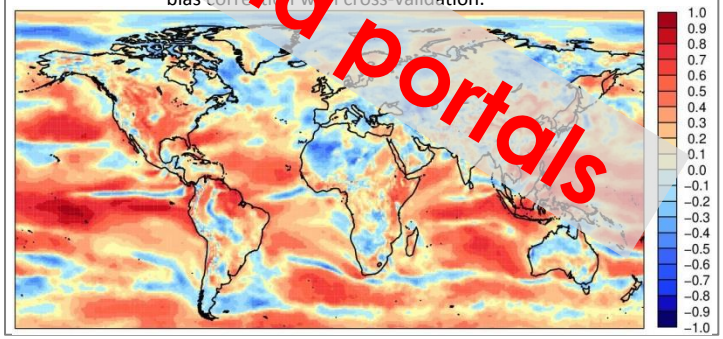
- Supports datasets stored locally or in ESGF (OPeNDAP) servers.
- Exploits multi-core capabilities
- API available
- Collects observational and experimental datasets stored in multiple conventions:
  - NetCDF3, NetCDF4
  - Supports specific folder and file naming conventions.

BASIC STATISTICS

SCORES
EXTREMES

PLOTS

Anomaly Correlation Coefficient. 10M Wind Speed ECMWF S4 1 month lead with start dates once a year on first of November and Era-Interim in DJF from 1981 to 2011. Simple bias correction with cross-validation.



Case 1: Data streaming for air quality forecasting

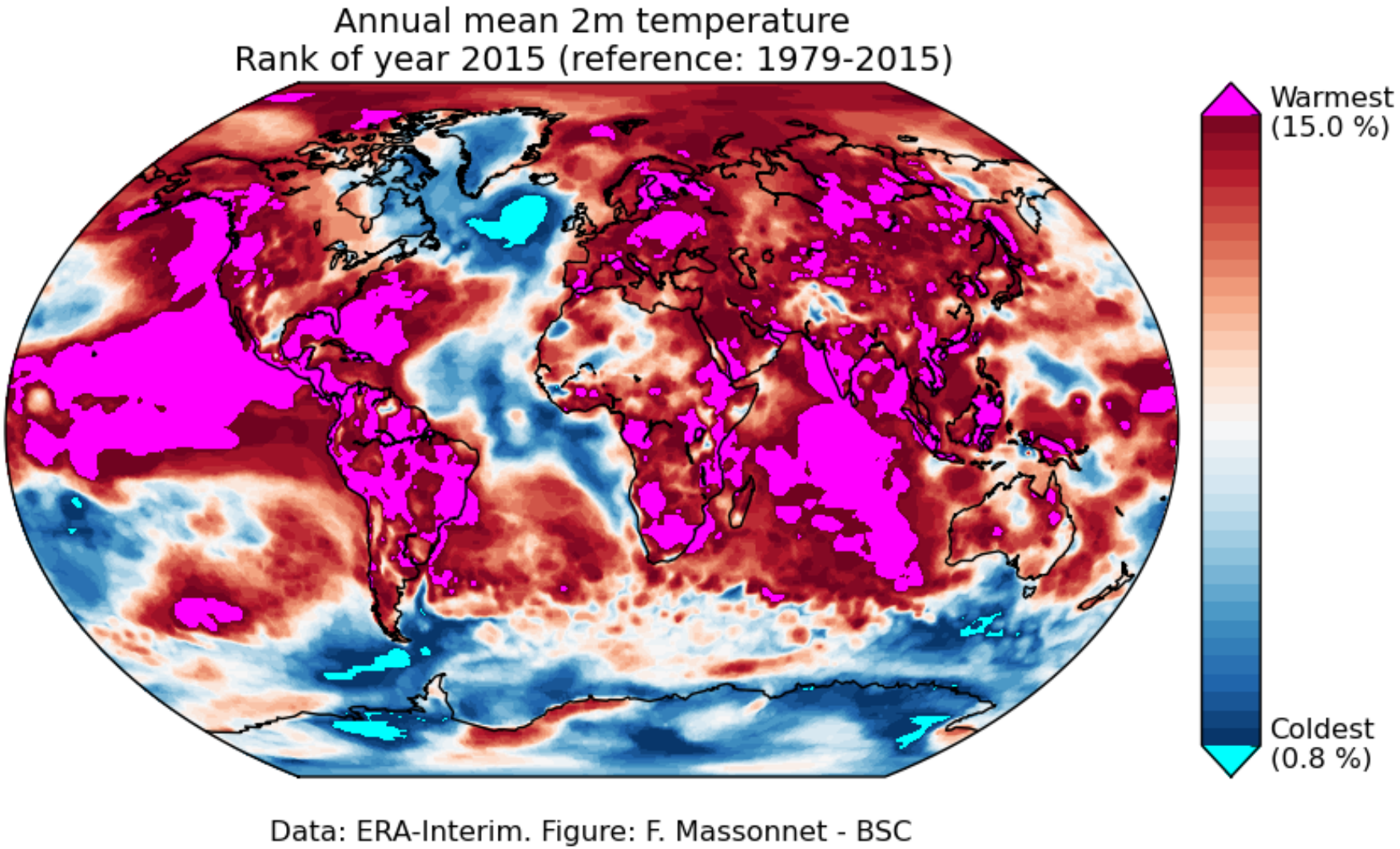
Case 2: Simultaneous analytics and HPC in climate prediction

Case 3: Analytics as a service

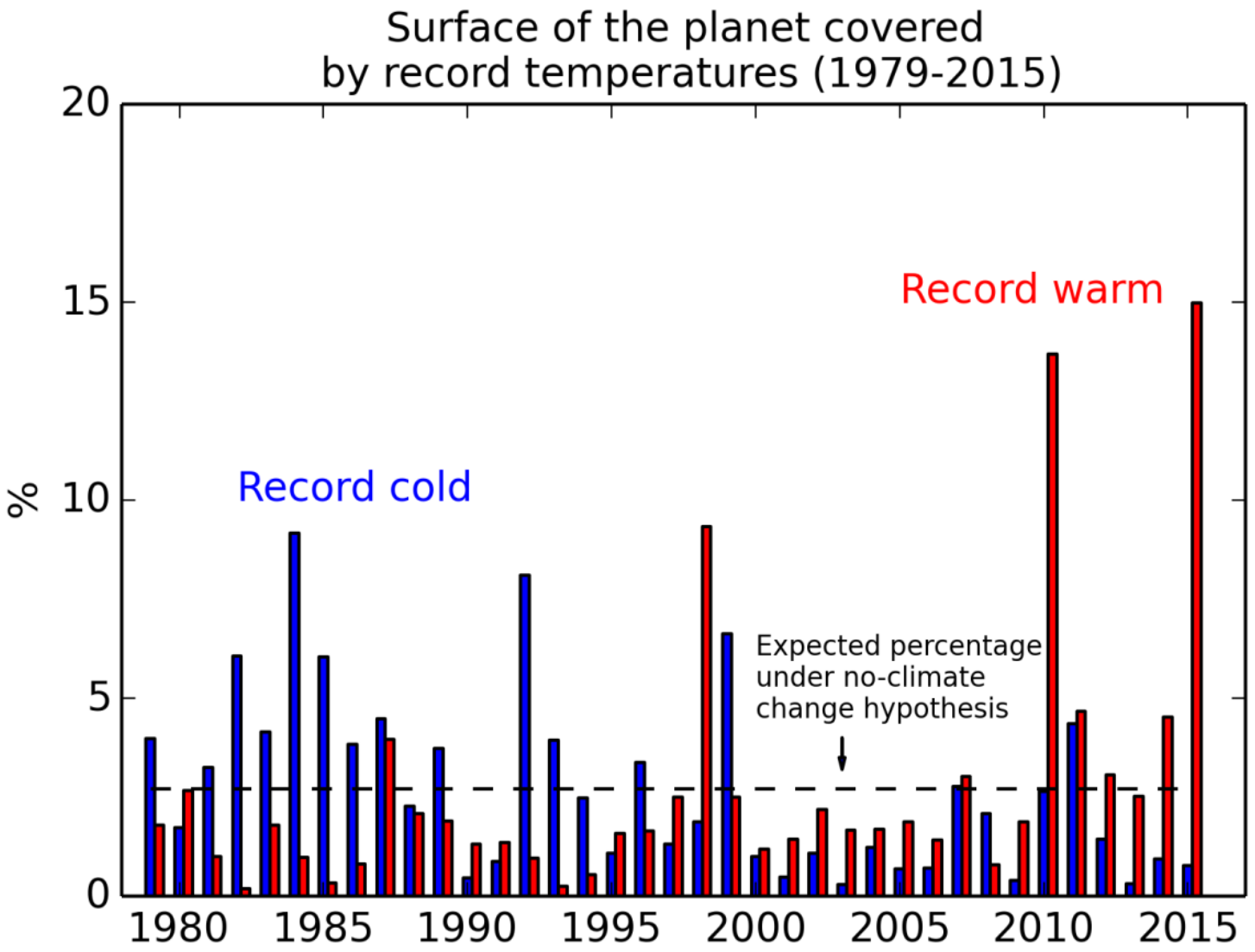
Climate change is taking place

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA

Rank of the 2015 annual mean temperature over the last 37 years from ERA Interim.



Climate change is taking place



F. Massonnet (BSC)

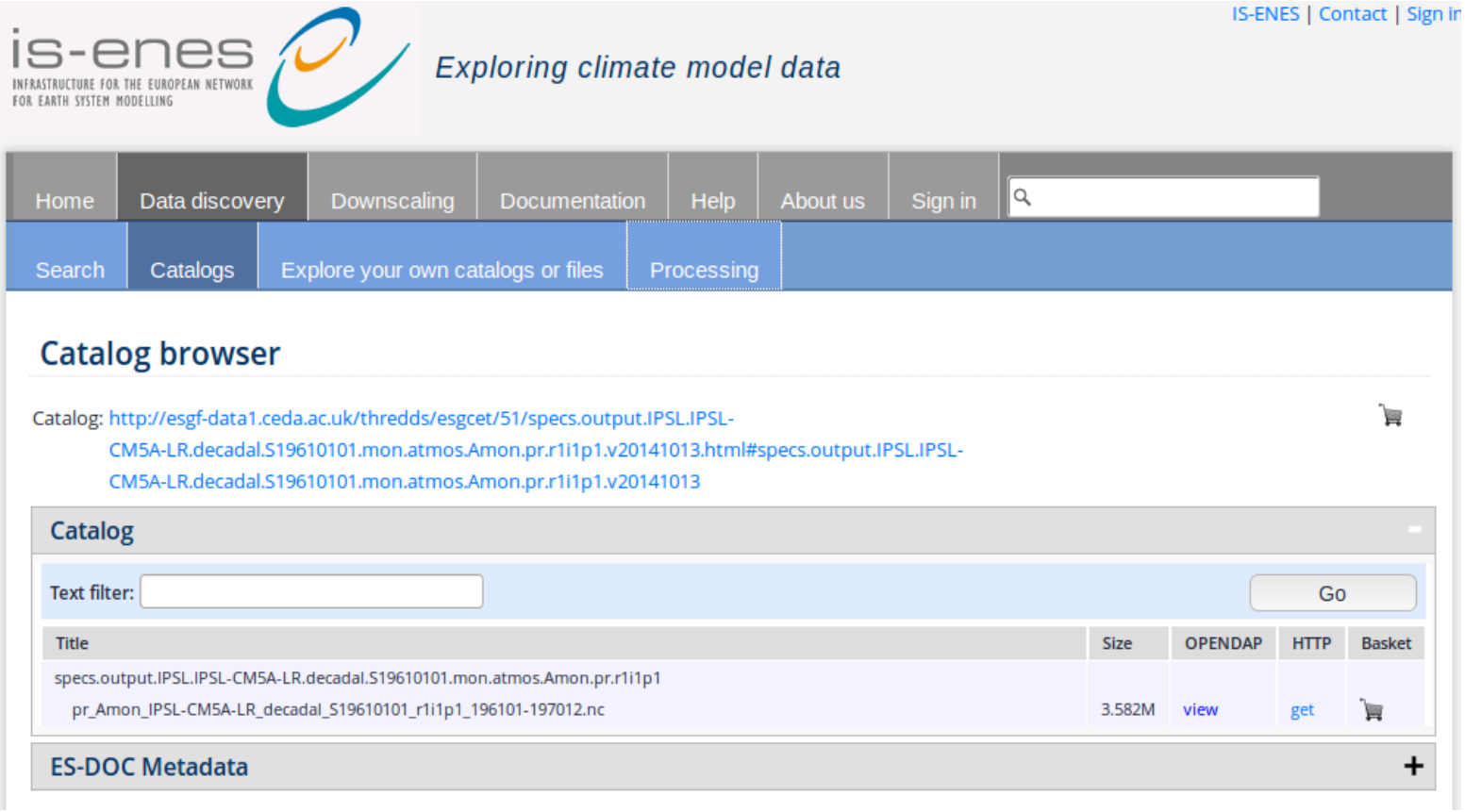
Downstream services

Users need services to respond to their questions.

An example of downstream service http://climate4impact.eu.

This is a key aspect for the success of our research.

Based on access to **ESGF**.

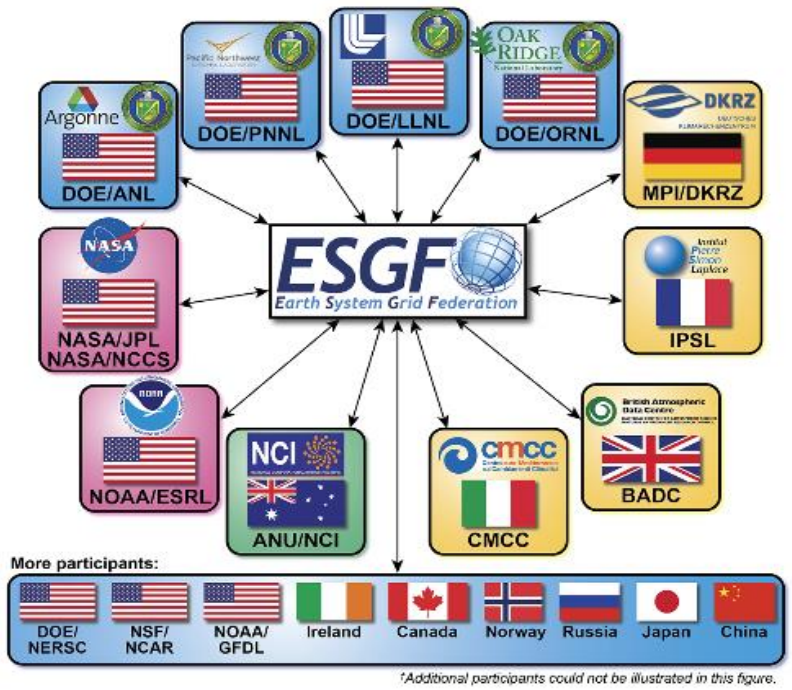


Data access of downstream services

The Earth System Grid Federation is an open source effort providing a robust, distributed data platform, enabling worldwide access to peta/exascale weather and climate data.

ESGF promotes common formatting, has discovery tools and data indexing.

EUDAT, a pan-European network, is working on general-purpose additional solutions.



ESGF is organized as a framework of worldwide **distributed nodes**

- **Data nodes** (THREDDS server, gridFTP server, …)
- **Index nodes** (database description)
- **Identity nodes** (authentication)
- **Compute nodes** (data analysis and visualization)

Need to compute where the data are

ESGF distributes mainly CMIP (Coupled Model Intercomparison Project) data. CMIP is one of the bases of the assessment reports of the Intergovernmental Panel on Climate Change (IPCC).

	CMIP1 (1996)	CMIP2 (1997)	CMIP3 (2005)	CMIP5 (2010)
Number of experiments	1	2	12	110
Institutions participating	16	18	15	24
Number of models	19	24	21	50
Total dataset size	1 GB	540 GB	36 TB	3.3 PB

Derived variables require either huge data traffic or large computing power and memory

100 times increase from one CMIP and the next one

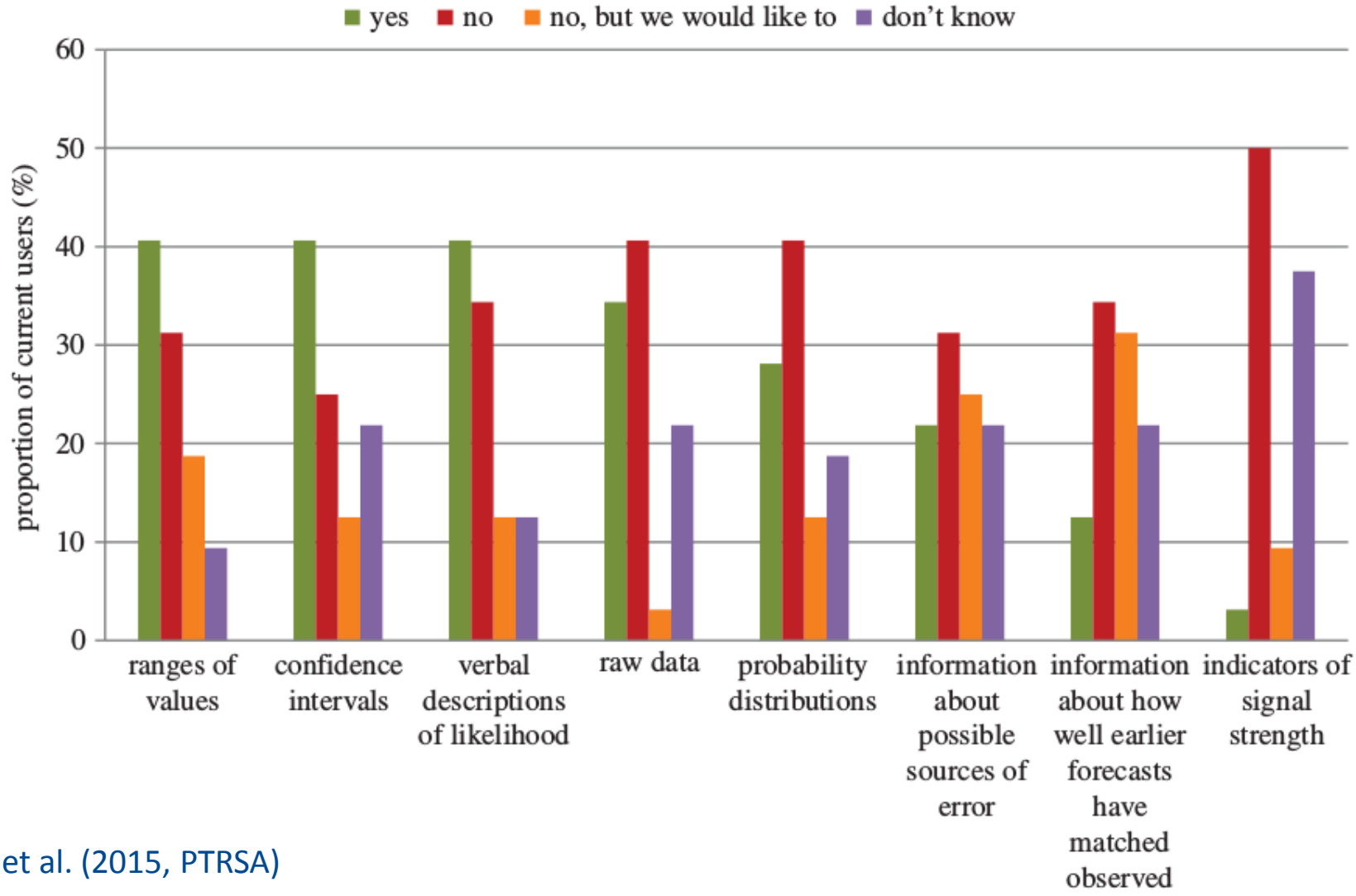
Service-driven research

Climate data is not climate information

Ethical Framework for Climate Services four core elements: integrity, transparency, humility and collaboration.

Climate services research

Proportion of users of seasonal forecasts (n = 32) indicating whether they received different forms of information about uncertainty.



Taylor et al. (2015, PTRSA)

National and international initiatives

- Spanish Network for Big Data in weather, climate and air quality.
  - Initiative started with a workshop in Madrid the 13th of November 2015.
  - Foster the discussion about solutions to issues linked to Big Data in our sectors among the research, operational and industrial Spanish community
  - http://www.bsc.es/projects/earthscience/bigdata
- Research Data Alliance (international initiative with the objective to promote data sharing among all kinds of research communities) Interest Group in weather, climate and air quality
  - Workshop held in Barcelona the 11th of February 2016
  - http://rd-alliance.org

Some more open questions

- **Education**: in the era of open data, the community could take advantage of the open education opportunities.
- **Heterogeneity**: how to link and merge our data to those from communities with larger impact (urban development, arts, social responsibility)?
- **Technology**: how can we make the most of a rapidly evolving technology (heterogeneous nodes, Big Data software, mobile data capture, storage/compression, outsourcing)?
- **Awareness**: what are the priority issues for our communities and how can we work together?
- **Industry engagement:** how involving the private sector to participate in and benefit from the discussion?